4. 로지스틱 회귀

5) 진단

5. 진단

- 진단
  1. 모형진단: 모형의 가정만족여부 확인
  2. 관찰값 진단: 특이한 관찰값(이상값 혹은 영향력이 큰 관찰값 탐지)

- 모형진단: link function(logit, probit 등)의 적절성 혹은 변수의 변환 필요성 등을 진단하는 단계

- 관찰값 진단
  - 이상값 탐지: 잔차, leverage 등
  - 영향력이 큰 관찰값 탐지: Cook 통계량, Dfbeta 등

1) 잔차

- 선형회귀모형에서의 잔차: $e_i = Y_i - \hat{Y}_i$
  - 오차항 ε_i에 대응되는 통계량
  - 로지스틱 회귀모형에서는 의미가 없음

- 로지스틱 회귀모형에 적합한 잔차
  - Pearson 잔차
  - Deviance 잔차

- Pearson 잔차(r_P)

  - 정의:

$$r_{P_i} = \frac{Y_i - \hat{\pi}(\boldsymbol{x}_i)}{\sqrt{\hat{\pi}(\boldsymbol{x}_i)(1 - \hat{\pi}(\boldsymbol{x}_i))}}$$

  - 명칭 유래: $\sum_{i=1}^{n} r_{P_i}^2$이 모형의 적합도를 평가하기 위한 Pearson 카이제곱 통계량과 같아지기 때문

  - r_{P_i}가 큰 값을 갖게 되면 Y_i가 모형의 적합도에 나쁜 영향을 주는 관찰값이라는 것을 의미

  - 큰 값의 기준: 명확하지 않음(교재 167쪽 내용 적용에 주의)
  - (x_i, r_{P_i})의 산점도 보다는 (i, r_{P_i})의 산점도가 더 효과적

- Deviance 잔차(r_D)
  - 정의:

$$r_{D_i} = \text{sign}(Y_i - \hat{\pi}_i)\sqrt{-2[Y_i \log \hat{\pi}_i + (1 - Y_i)\log(1 - \hat{\pi}_i)]}$$

단, $$\text{sign}(Y_i - \hat{\pi}_i) = \begin{cases} +1, & \text{if } Y_i \geq \hat{\pi}_i \\ -1, & \text{if } Y_i < \hat{\pi}_i \end{cases}$$

$$\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i)$$

  - 명칭 유래: $\sum_{i=1}^{n} r_{D_i}^2$이 현재 모형의 deviance인 통계량 D가 되기 때문
  - r_{D_i}가 큰 값을 갖게 되면 Y_i가 모형의 적합도에 나쁜 영향을 주는 관찰값 이라는 것을 의미

2) 표준화 잔차

- Pearson 잔차와 deviance 잔차들의 분산은 1보다 작음
- 잔차의 분산이 대략 1이 되도록 조정할 필요가 있음

- 표준화 과정
  - Pearson 잔차: 잔차 $Y_i - \hat{\pi}_i$을 자신의 표본오차로 나눈다
  - deviance 잔차: r_{D_i}를 자신의 표본오차로 나눈다

- 잔차의 표본오차에는 모자행렬(Hat matrix)의 대각원소인 leverage, h_{ii}가 포함됨.

- Leverage
  - 선형회귀모형
    - Hat matrix, H: $\hat{Y} = HY$
    - leverage, h_{ii}: 모자행렬 H의 대각원소
    - h_{ii}가 큰 값을 갖게 되면 i번째 관찰값이 모형의 적합에 큰 영향을 줄 가능성이 높다는 것을 의미
  - 로지스틱 회귀모형
    - Hat matrix: 선형회귀모형과는 다르게 정의됨
    - leverage의 정의: 교재 164쪽 식 4.4 참조
    - 의미: 선형회귀모형의 경우와 유사하지만, $\hat{\pi}(x_i) \approx 0$ 혹은 $\hat{\pi}(x_i) \approx 1$이 되는 경우에는 자료의 중심에 있지 않더라도 작은 값을 가질 수 있음

- 표준화 Pearson 잔차(Standardized Pearson residual)

$$r_{PS_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}} = \frac{r_{P_i}}{\sqrt{1-h_{ii}}}$$

- 표준화 deviance 잔차(Standardized deviance residual)

$$r_{DS_i} = \frac{r_{D_i}}{\sqrt{1-h_{ii}}}$$

3) 영향력이 큰 관찰값 탐지

- 회귀분석 결과에 큰 영향을 미치는 관찰값 탐색이 목적
- 기본 개념: 특정 관찰값을 포함한 분석결과와 포함하지 않은 분석결과의 비교
- 많이 사용되는 통계량
  - DFbeta: 개별 회귀계수에 대한 영향력 탐지
  - Cook's distance: 회귀계수 벡터에 대한 영향력 탐지

- DFbeta

$$\text{DFbeta}_{j(i)} = \frac{\Delta\hat{\beta}_{j(i)}}{\sqrt{\widehat{Var}(\hat{\beta}_j)}}, \qquad j = 1, \cdots, p \quad i = 1, \cdots, n$$

$$\Delta\hat{\beta}_{j(i)} = \hat{\beta}_j - \hat{\beta}_{j(i)}$$

$\hat{\beta}_j$: 모든 자료를 사용하여 구한 β_j 추정값
$\hat{\beta}_{j(i)}$: i번째 자료를 제외하고 구한 β_j 추정값

$\text{DFbeta}_{j(i)}$의 절대값이 크다면 i번째 관찰값이 β_j의 추정에 큰 영향을 미쳤다고 할 수 있음

- Cook's distance

  - Cook's distance는 i번째 관찰값을 포함시켰을 때와 제외시켰을 때 회귀계수 추정량 벡터 $\hat{\boldsymbol{\beta}}$의 차이를 표준화한 통계량

  - GLM에서 사용되는 Cook's distance

$$C_i = \frac{r^2{}_{PSi}}{p+1} \times \frac{h_{ii}}{1-h_{ii}}, \qquad i = 1, \cdots, n$$

  - Cook's distance가 크다면 i번째 관찰값이 회귀계수 추정에 큰 영향을 미쳤다고 볼 수 있음

4) R에서 진단 실시

- 관련된 통계량 계산 및 출력
  - (표준화)잔차: residuals( ), rstandard( )
  - Dfbeta, leverage, Cook's distance: dfbetas( ), hatvalues( ), cook.distance( )

- 적절한 그래프 작성(패키지 car)
  - 잔차 그래프: residualPlots( )
  - 영향력이 큰 관찰값 탐지: dfbetasPlots( ), inflndexPlot( ), influencePlot( )

예제 4.2

```
> library(carData)
> fit <- glm(lfp ~ . -k618 -hc, family=binomial, Mroz)
```

(1) 잔차

```
> r1 <- residuals(fit, type="pearson")
> r2 <- residuals(fit, type="deviance")    # default

> cbind(pearson=r1[1:5], deviance=r2[1:5])
    pearson  deviance
1 1.0126913 1.1881399
2 0.7076713 0.9011076
3 1.0262008 1.1994541
4 0.6908877 0.8834406
5 0.6738134 0.8652575
```

(2) 표준화 잔차

```
> r3 <- rstandard(fit, type="pearson")
> r4 <- rstandard(fit, type="deviance")  # default

> cbind(pearson=r3[1:5], deviance=r4[1:5])
    pearson  deviance
1 1.0169670 1.1931563
2 0.7112617 0.9056793
3 1.0308468 1.2048844
4 0.6943320 0.8878447
5 0.6772211 0.8696333
```

(3) leverage

```
> hatvalues(fit)[1:5]
          1           2           3           4           5
0.008391066 0.010070156 0.008993577 0.009896392 0.010038247
```

(4) Dfbeta(표준화 Dfbeta)

```
> dfbetas(fit)[1:5,]
  (Intercept)           k5          age        wcyes          lwg
1 0.031590700  0.04953680 -0.02891614 -0.029436766  0.02262410
2 0.073073072 -0.04005855 -0.06329411 -0.005676523 -0.03373092
3 0.009013129  0.05837133 -0.01220317 -0.037021799  0.04624895
4 0.065695847 -0.03258143 -0.04227339  0.008323221 -0.04420896
5 0.023153840  0.02061330 -0.02714774  0.044369803  0.01411801
           inc
1 -0.023628204
2  0.008629781
3 -0.021681538
4 -0.028481593
5 -0.013718853
```

(5) Cook's distance

```
> cooks.distance(fit)[1:5]
           1            2            3            4            5
0.0014586102 0.0008577077 0.0016072854 0.0008031179 0.0007750846
```

- 효과적인 진단을 위해서는 적절한 그래프 작성이 필수적

- 패키지 car에 유용한 함수 다수 존재

    1) residualPlots( ) : residual plots 작성 및 curvature test 실시

    2) dfbetasPlots( ) : 표준화 Dfbeta의 index plot 작성

    3) infIndexPlot( ) : Cook's distance, Studentized residual, leverage들의 개별 index plot 작성

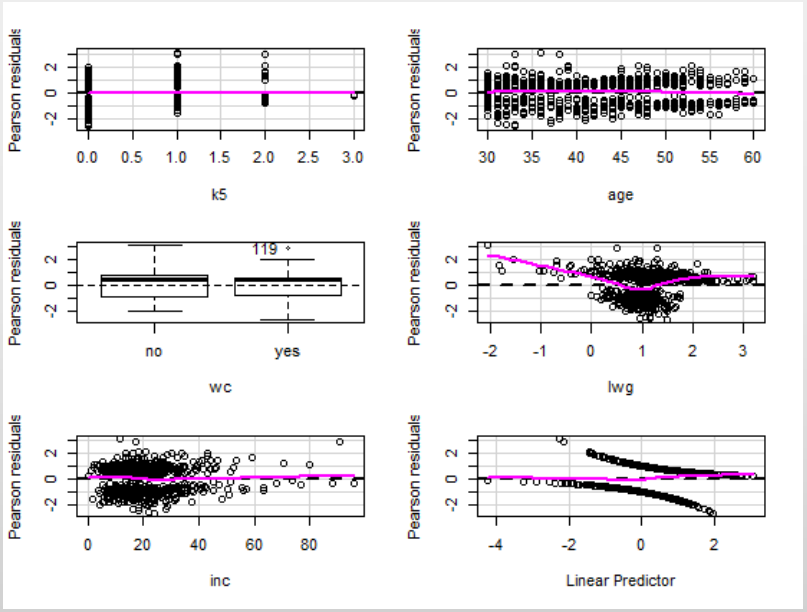
    4) influencePlot( ) : Studentized residual과 leverage의 산점도를 기반으로 Cook's distance 크기에 의해 작성된 bubble plot

(6) 모형 fit의 잔차 산점도 작성 및 curvature test 실시

```
> library(car)

> residualPlots(fit)
    Test stat Pr(>|Test stat|)
k5     0.1102           0.73990
age    0.6284           0.42793
wc
lwg  152.7522           < 2e-16 ***
inc    3.2493           0.07145 .
```

curvature test:
- 선형 관계 여부 확인
- 모형에 lwg^2을 포함시킬 필요가 있는 것으로 보임



잔차 산점도

- lwg^2 모형에 추가

```
> fit.1 <- update(fit, . ~ . + I(lwg^2))
Warning message:
glm.fit: 적합된 확률값들이 0 또는 1 입니다
> summary(fit.1)
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.518168   0.821533   7.934 2.12e-15 ***
k5          -1.527347   0.223132  -6.845 7.65e-12 ***
age         -0.068245   0.012794  -5.334 9.61e-08 ***
wcyes        0.139512   0.239424   0.583 0.560097
lwg         -7.763915   1.094941  -7.091 1.33e-12 ***
inc         -0.033799   0.008864  -3.813 0.000137 ***
I(lwg^2)     4.580429   0.566054   8.092 5.88e-16 ***
```

- 모형에 추가된 lwg^2은 유의적
- 기존 변수인 wc가 비유의적이 됨

- 모형 선택: fit vs fit.1

```
> fit$aic
[1] 918.4554
> fit.1$aic
[1] 767.7032
```

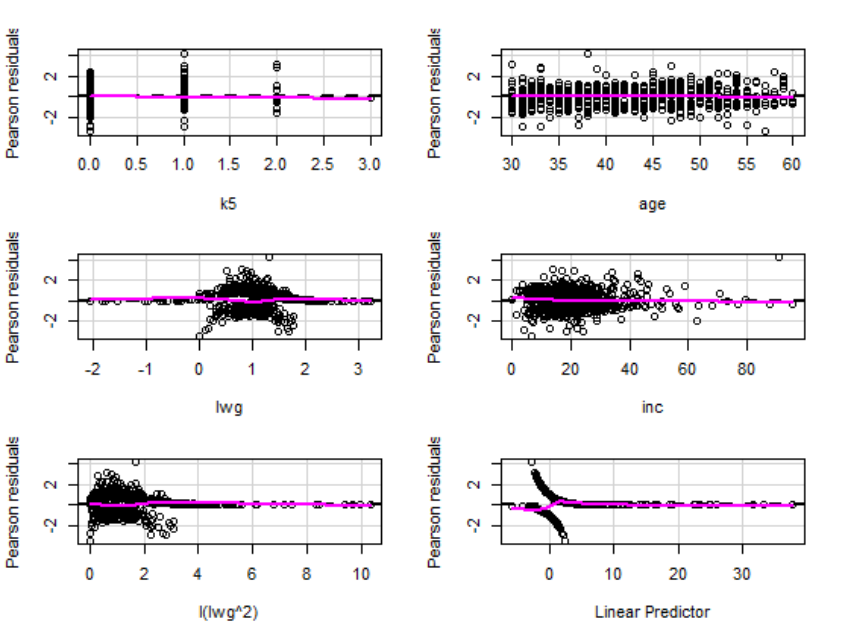
- AIC 값이 훨씬 작은 fit.1을 선택하는 것이 좋을 듯
- 모형 fit.1에서 비유의적인 변수 wc는 어떻게?

```
> fit.2 <- update(fit.1, . ~ . - wc)
> fit.2$aic
[1] 766.0427
```

- AIC 값은 큰 차이 없음
- 부인 학력이 반드시 필요한 변수가 아니라면 제거하는 것이 좋을 듯

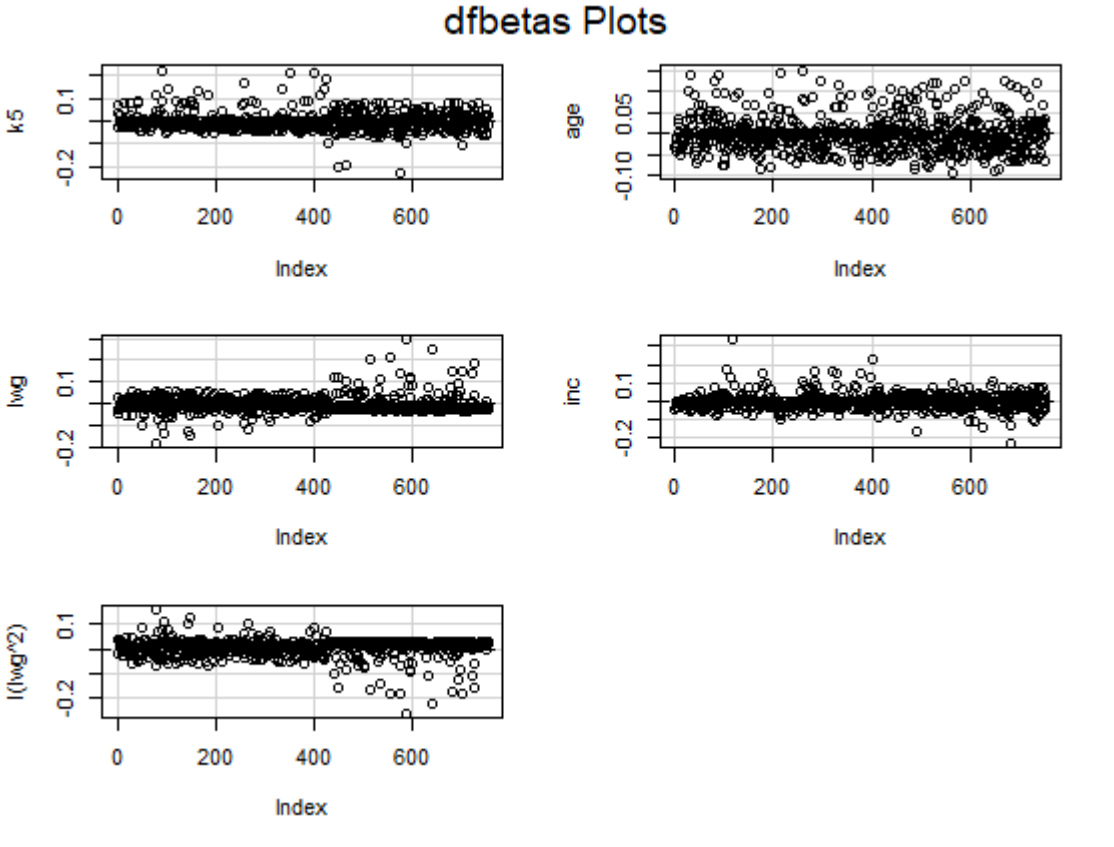
(6-1) 모형 fit.2의 잔차 산점도

```
> residualPlots(fit.2)
           Test stat Pr(>|Test stat|)
k5            0.2215             0.6379
age           1.0907             0.2963
lwg           0.0000             1.0000
inc           2.3168             0.1280
I(lwg^2)      0.1372             0.7111
```



(7) 모형 fit.2의 Dfbetas plot

```
> dfbetasPlots(fit.2)
```



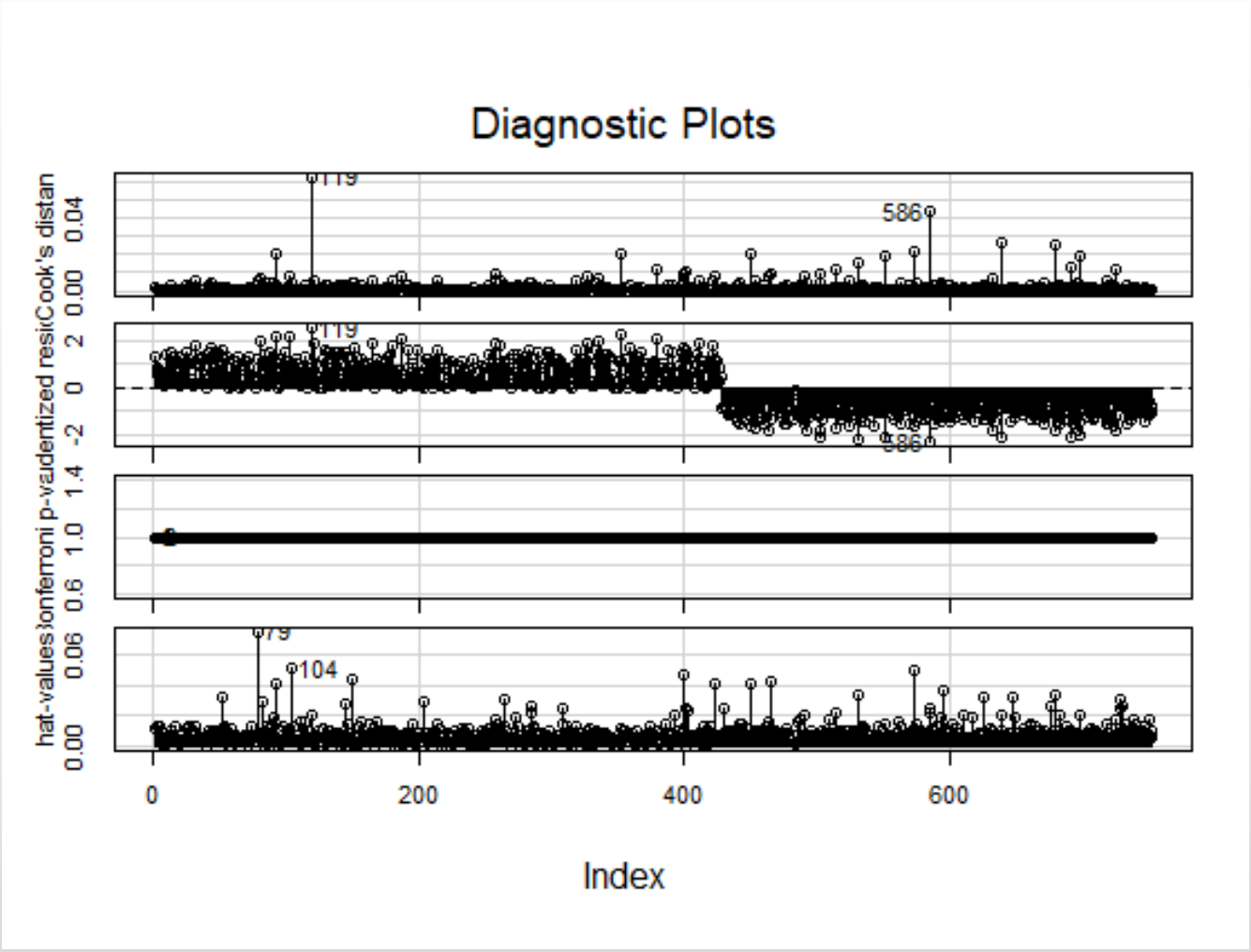
(8) 모형 fit.2의 Influence Index Plot

```
> infIndexPlot(fit.2)
```

- Cook's distance, Studentized residuals, outlier test에 대한 Bonferroni p-value, leverage의 index plot

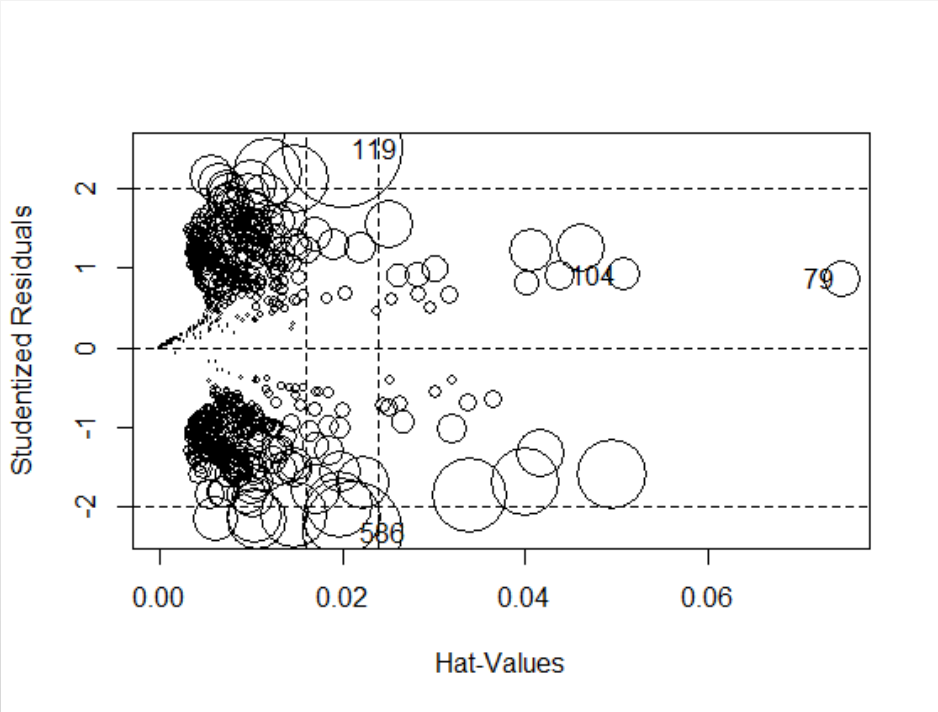
- Studentized residual:
  - 선형회귀모형: i번째 자료를 제외하고 나머지 자료로 추정한 Y_i의 적합값과 Y_i의 차이.
  - 로지스틱 회귀모형: 표준화 pearson 잔차와 표준화 deviance 잔차를 사용하여 근사값을 구함. 함수 rstudent( )로 계산.

- outlier test: studentized residual의 절대값이 가장 큰 관찰값에 대한 outlier test의 Bonferroni p-value.


Diagnostic Plots
Cook's distan
0.04
0.00
119
586
Standardized resi
2
0
-2
119
586
Bonferroni p-value
1.4
1.0
0.6
hat-values
0.06
0.00
79
104
0
200
400
600
Index

(9) 모형 fit.2의 Influence Plot

```
> influencePlot(fit.2)
       StudRes         Hat        CookD
79   0.8546115 0.07484018 0.006072504
104  0.9149417 0.05082217 0.004696899
119  2.4991933 0.02012581 0.062533536
586 -2.3176072 0.02107789 0.043637000
```



Y축: Studentized residual
X축: leverage
점의 지름: Cook's distance에 비례